AD

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 December 2000 (14.12.2000)

PCT

(10) International Publication Number
WO 00/75346 A1

(51) International Patent Classification[7]: C12N 15/62, 15/70, 1/21, C12P 21/02, C07K 14/47, 14/245, 1/113 // (C12N 1/21, C12R 1:19)

(21) International Application Number: PCT/GB00/01981

(22) International Filing Date: 23 May 2000 (23.05.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
9913437.1 9 June 1999 (09.06.1999) GB

(71) Applicant *(for all designated States except US)*: MEDICAL RESEARCH COUNCIL [GB/GB]; 20 Park Crescent, London W1N 4AL (GB).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: FERSHT, Alan [GB/GB]; MRC Unit for Protein Function and Design, Department of Chemistry, Lensfield Road, Cambridge CB2 1EW (GB).

(74) Agent: MASCHIO, Antonio; D Young & Co, 21 New Fetter Lane, London EC4A 1DA (GB).

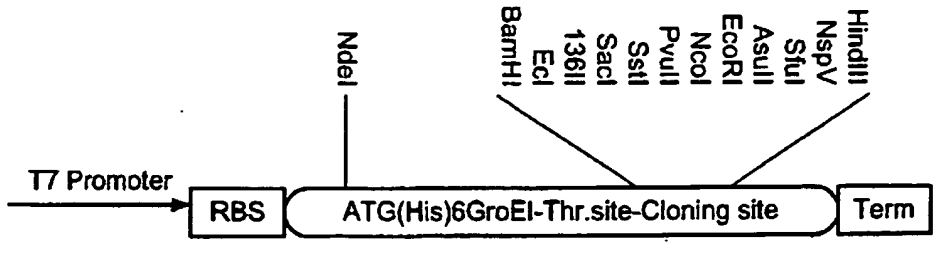
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *With international search report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: FUSION PROTEINS COMPRISING A FRAGMENT OF A CHAPERON POLYPEPTIDE



(57) Abstract: The invention relates to a fusion protein comprising: a) a first region comprising a fragment of a chaperone polypeptide; and b) a second region not naturally associated with the first region comprising a polypeptide sequence of interest, as well as to nucleic acid sequences encoding such a fusion protein and methods for expressing polypeptides based on such nucleic acids.

WO 00/75346 A1

FUSION PROTEINS COMPRISING A FRAGMENT OF A CHAPERON POLYPEPTIDE

The present invention relates to chaperone polypeptides which are active in the folding and maintenance of structural integrity of other proteins and the use thereof as fusion partners to assist in the expression of polypeptides in expression systems. The invention also relates to nucleic acids encoding chaperone polypeptides and fusion proteins as described, vectors comprising these nucleic acids, and host cells modified with the nucleic acids or vectors so as to express the fusion protein(s).

Chaperones are in general known to be large multisubunit protein assemblies essential in mediating polypeptide chain folding in a variety of cellular compartments. Families of chaperones have been identified, for example the chaperonin hsp60 family otherwise known as the cpn60 class of proteins are expressed constitutively and there are examples to be found in the bacterial cytoplasm (GroEL), in endosymbiotically derived mitochondria (hsp60) and in chloroplasts (Rubisco binding protein). Another chaperone family is designated TF55/TCP1 and found in the thermophilic archaea and the evolutionarily connected eukaryotic cytosol. A comparison of amino acid sequence data has shown that there is at least 50% sequence identity between chaperones found in prokaryotes, mitochondria and chloroplasts (Ellis R J and Van der Vies S M (1991) Ann Rev Biochem 60: 321-347).

A typical chaperonin is GroEL which is a member of the hsp60 family of heat shock proteins. GroEL is a tetradecamer wherein each monomeric subunit (cpn60m) has a molecular weight of approximately 57kD. The tetradecamer facilitates the *in vitro* folding of a number of proteins which would otherwise misfold or aggregate and precipitate. The structure of GroEL from *E. coli* has been established through X-ray crystallographic studies as reported by Braig K *et al* (1994) Nature 371: 578-586. The holo protein is cylindrical, consisting of two seven-membered rings that form a large central cavity which according to Ellis R J and Hartl F U (1996) FASEB Journal 10: 20-26 is generally considered to be essential for activity. Some small proteins have been demonstrated to fold from their denatured states when bound to GroEL (Gray T E and Fersht A R (1993) J

Mol Biol 232: 1197-1207; Hunt J F *et al* (1996) Nature 379: 37-45; Weissman J S *et al* (1996) Cell 84: 481-490; Mayhew M *et al* (1996) Nature 379: 420-426; Corrales F J and Fersht A R (1995) Proc Nat Acad Sci 92: 5326-5330) and it has been argued that a cage-like structure is necessary to sequester partly folded or assembled proteins (Ellis R J and Hartl F U (1996) *supra*.

The entire amino acid sequence of *E. coli* GroEL is also known (see Braig K *et al* (1994) *supra*) and three domains have been ascribed to each cpn60m of the holo chaperonin (tetradecamer). These are the intermediate (amino acid residues 1-5, 134-190, 377-408 and 524-548), equatorial (residues 6-133 and 409-523) and apical (residues 191-376) domains.

Monomers of GroEL have been induced by urea or pressure, but they are inactive and have to reassociate to form the central cavity in order to facilitate the refolding of rhodanese (Mendoza J A *et al* (1994) J Biol Chem 269: 2447-2451; Ybarra J and Horowitz P M (1995) J Biol Chem 270: 22962-22967).

GroEL facilitates the folding of a number of proteins by two mechanisms; (1) it prevents aggregation by binding to partly folded proteins (Goloubinoff P *et al* (1989) Nature 342: 884-889; Zahn R and Plückthun A (1992) Biochemistry 31: 3249-3255), which then refold on GroEL to a native-like state (Zahn R and Plückthun A (1992) Biochemistry 31: 3249-3255; Gray T E and Fersht A R (1993) J Mol Biol 232: 1197-1207); and (2) it continuously anneals misfolded proteins by unfolding them to a state from which refolding can start again (Zahn R *et al* (1996) Science 271: 642-645). Some mutations in the apical domain led to a decrease in polypeptide binding (Fenton W A *et al* (1994) Nature 371: 614-619), suggesting that this domain is involved in the binding of polypeptides. Electron microscopy suggests that denatured protein binds to the inner side of the apical end of the GroEL-cylinder (Chen S *et al* (1994) Nature 371: 261-264). The equatorial domain has been shown from the 2.4 Å crystal structure of ATPγS-ligated GroEL (Boisvert D C *et al* (1996) Nature Structure Biology 3: 170-177) and mutagenesis studies (Fenton W A *et al* (1994) Nature 371: 614-619) to have the nucleotide binding sites. Binding and hydrolysis of ATP is cooperative (Bochkareva E S *et al* (1992) J Biol

Chem 267: 6796-6800; Gray T E and Fersht A R (1991) FEBS Lett 292: 254-258), and lowers the affinity for polypeptides (Jackson G S *et al* (1993) Biochemistry 32: 2554-2563). Most of the intermolecular contacts between the subunits of GroEL are between the equatorial domain. The intermediate domain connects the other two domains, transmitting allosteric effects (Braig K *et al* (1994) Nature 371: 578-586; Braig K *et al* (1995) Nature Struct Biol 2: 1083-1094).

The crystal structure of GroEL shows unusually high *B-factors* for the apical domain compared with the equatorial or intermediate domain, and the *B-factors* vary considerably within the domain (Braig K *et al* (1994) Nature 371: 578-586; Braig K *et al* (1995) Nature Struct Biol 2: 1083-1094; Boisvert D C *et al* (1996) Nature Structure Biology 3: 170-177). The high overall *B-factor* seems to result from a static disorder within the asymmetric unit and probably throughout the crystals of GroEL, and has been attributed to rigid-body movements generated by hinge-like β-sheets in the intermediate domain. Regions of high flexibility have also been observed in the 2.8Å structure of the co-chaperonin GroES (Hunt J F *et al* (1996) Nature 379: 37-45). A mobile loop has been shown to be directly involved in ADP-dependent binding to the apical domain (Landry S J *et al* (1993) Nature 364: 255-258). Binding of GroES leads to a conformational change of GroEL and a concomitant enlargement of the GroEL-cavity (Chen S *et al* (1994) Nature 371: 261-264), in which the encapsulated polypeptide substrate can refold to a native-like state without the danger of aggregation (Martin J *et al* (1993) Nature 366: 228-233; Weissman J S *et al* (1995) Cell 83: 577-587).

Monomeric forms of GroEL have been induced by site-directed mutagenesis and expressed and although these bind to rhodanese they do not affect its refolding (White Z W *et al* (1995) J Biol Chem 270: 20404-20409).

Yoshida *et al* (1993) FEBS 336: 363-367 report that a 34kD proteolytic fragment of *E. coli* GroEL which lacks 149 NH_2-terminal residues and ~93 COOH-terminal residues (GroEL 150-456) facilitates refolding of denatured rhodanese in the absence of GroES and ATP. Although the proteolytic fragment GroEL 150-456 elutes as a monomer during gel filtration, it still comprises the apical domain and significant portions of the

intermediate and equatorial domains, the latter of which determine the intersubunit contacts of GroEL (Braig K *et al* (1994) *supra*), thus allowing transient formation of the central cavity thereby accounting for the chaperonin activity which is observed.

In any event, the mode of rhodanese refolding by GroEL 150-456 is very different from that brought about by the holo protein; the yield of productive refolding is low, folding is rapidly saturated with time, and it is not affected by GroES and ATP. Efficient release and folding requires the hydrolysis of ATP (Landry S J *et al* (1992) Nature 355: 455-457; Gray T E and Fersht A R (1992) FEBS Lett 282: 254-258; Jackson G S *et al* (1993) Biochemistry 32: 2554-2563; Todd M *et al* (1993) Biochemistry 32: 8560-8567.)

EP-A-0 650 975 (NIPPON OIL CO LTD) discloses chaperonin molecules and a method of refolding denatured proteins using GroEL chaperonin 60 monomers (cpn60m) obtained from *Thermus thermophilus*. The holo-chaperonin was first extracted and then purified from the bacterial source according to the method of Taguchi *et al* (1991) J Biol Chem 266: 22411-22418. The cpn60m was then produced by treatment of the holo-chaperonin with trifluoroacetic acid (TFA) followed by reverse phase (rp) HPLC of the resulting denatured protein. A peak fraction containing the approximately 57kD cpn60m was obtained. The refolding activity of the cpn60m was assayed in solution by monitoring the regain in activity of inactivated rhodanese, which in specific activity terms amounted to about only 25% of the specific activity of the rhodanese prior to inactivation. When background spontaneous rhodanese refolding is subtracted then there is only an approximately 20% refolding activity.

As well as cpn60m, EP-A-0 650 975 also discloses the use of an approximately 50kD N-terminal deletion fragment of cpn60m wherein the N-terminal amino acid residues up to (but not including) the Thr residue at position 79 are removed by proteolysis. This 50kD fragment showed an approximately 35% (about 30% when background is subtracted) rhodanese refolding activity when in solution.

Taguchi H *et al* (1994) J Biol Chem 269: 8529-8534 is a scientific report on which the invention of EP-A-0 650 975 is based. A transiently formed GroEL tetradecamer (the

holo-chaperonin) was perceived to exist when the chaperonin monomers are present in solution. Consequently, the refolding activity of these preparations can be seen to be caused by the presence of holo chaperonin, not monomers. To test this, Taguchi *et al* immobilised cpn60m to a chromatographic resin to exclude the possibility of holo chaperonin formation. When immobilised and therefore when in truly monomeric form, cpn60m exhibited only about 10% rhodanese refolding activity.

The refolding of rhodanese has been a common and convenient assay to determine chaperonin activity but it has been observed that there are significant problems with the assay which cast serious doubt on existing assertions of refolding activity based on this assay. The fact is that rhodanese refolds spontaneously in the absence of molecular chaperones with the yield of refolded rhodanese increasing progressively as the rhodanese concentration decreases (see Taguchi *et al* (1994) *supra*). The 10% of rhodanese refolding activity reported in EP-A-0 650 975 for immobilised (truly monomeric) cpn60m is therefore too close to the spontaneous regain of activity by rhodanese to demonstrate that any monomeric chaperonin has a refolding activity towards proteins generally, let alone cpn60m and rhodanese.

Alconada A and Cuezva J M (1993) TIBS 18: 81-82 suggested that an "*internal fragment*" of GroEL may possess a chaperone activity on the basis of amino acid sequence similarity between the altered mRNA stability (ams) gene product (Ams) of *E. coli* and the central part of GroEL. The ams locus is a temperature-sensitive mutation that maps at 23 min on the *E. coli* chromosome and results in mRNA with an increased half-life. The ams gene has been cloned, expressed and shown to complement the ams mutation. The gene product is a 149-amino acid protein (Ams) with an apparent molecular weight of 17kD.

Chanda P K *et al* (1985) J Bacteriol 161: 446-449 found that a 17kD protein fragment corresponding to part of the L gene of the groE operon, when expressed in *E. coli* ams mutants restores the wild-type phenotype. This 17kD fragment was suggested as being an isolated, functional chaperonin protein module. The amino acid sequences of three chaperonins (*E. coli* GroEL, ribulose bisphosphate carboxylase (RUBPC) subunit-binding protein from *Triticum aestivum* and *Saccharomyces cerevisiae* mitochondrial hsp60) were

compared with the sequence of Ams. Residues 307-423 were found to correspond substantially between Ams and GroEL. These residues comprise nearly equivalent portions of both the intermediate and apical domains of GroEL.

The sequence alignments of Ams protein with the chaperonins noted above reveals a striking similarity (98%) between the amino-terminal four-fifths of Ams and the central part (approximately one-fifth) of *E. coli* GroEL chaperonin. The 50% sequence similarity between the Ams amino terminal region and the two other chaperonins is in line with the reported identity among the chaperonin family. The carboxy-terminal part of the Ams protein showed no similarity with chaperonins (<10% homology).

International Patent Application WO98/13496, the entire contents of which is incorporated herein by reference, describes fragments of chaperone molecules, termed minichaperones, which are effective in promoting the folding of unfolded or misfolded polypeptides. The fragments are monomeric in solution.

Recombinant DNA technology has allowed industry to produce many proteins of commercial importance. Proteins are produced in a wide variety of expression systems which are based on, for example, bacterial, yeast, insect, plant and mammalian cells. one of the problems associated with the production of proteins by recombinant means is that host cells contain enzymes which degrade proteins and the presence of such enzymes present particular difficulties in the production of small polypeptides. Moreover, polypeptides produced by recombinant DNA technology are frequently at least partially incorrectly folded, such that yields of biologically active molecules vary according to the ability of the expression system to promote correct folding. This can moreover be problematic in the production of polypeptides destined for chemical and physical analysis, for which structural homogeneity is highly relevant.

One approach to overcoming such difficulties is to express a recombinant protein of interest in the form of a fusion protein. DNA encoding the protein of interest is fused in-frame to a fusion partner protein and the resulting fusion is expressed. Often, a linker

sequence encoding a protease cleavage site between the two parts of the fusion is included to allow cleavage of the fusion after it has been recovered from its host cell.

The fusion partner protein is often one which may be recovered and purified by some form of highly specific affinity purification means. Examples of such proteins are well known in the art and include, for example, glutathione-S-transferase, maltose binding protein and β-lactamase.

However these fusion partner proteins are all relatively large and thus have a number of disadvantages. For example, it is essential to remove them before any meaningful procedure may be carried out on the protein of interest, since they are too large to enable it to function with any degree of independence. Many small polypeptides are still thus made by chemical synthesis.

Summary of the Invention

The present invention provides fusion proteins which incorporate chaperone fragments as fusion partners to promote high yield expression of correctly folded polypeptides in biological expression systems. It has been observed that consistently higher yields of recombinantly expressed polypeptides are obtained if the proteins are expressed as fusions with a chaperone fragment.

According to a first aspect of the invention, therefore, there is provided a fusion protein comprising:

a) a first region comprising a fragment of a chaperone polypeptide; and

b) a second region not naturally associated with the first region comprising a polypeptide sequence of interest.

The term "fusion protein" is used in accordance with its ordinary meaning in the art and refers to a single protein which is comprised of two or more regions which are derived

from different sources. Typically, a fusion protein is two proteins fused together by way of in-frame fusion of their respective nucleic acid coding sequences.

A "chaperone fragment", as referred to herein, is any fragment of a molecular chaperone which possesses the ability to promote the folding of a polypeptide *in vivo* or *in vitro*. Preferred fragments are described in International patent application WO98/13496, incorporated herein by reference. Especially preferred are fragments 191-375, 191-345 and 193-335 of GroEL. Advantageously, the GroEL is *E. coli* GroEL, as further described below.

The fusion protein according to the invention, in addition to the chaperone fragment, includes a desired polypeptide. The desired polypeptide is typically a polypeptide which it is desired to express by recombinant DNA techniques; it is expressed as a fusion with the chaperone fragment in order to increase the yield of correctly folded product, in accordance with the present invention. Many polypeptides may be expressed as fusion proteins according to the present invention. However, the expression of smaller polypeptides, up to about 250 amino acids in length, is preferred. Preferably, the polypeptides are between about 5 and about 100 amino acids in length.

Advantageously, the polypeptide is a eukaryotic polypeptide, such as a mammalian polypeptide.

Preferably, the fusion protein according to the invention comprises a cleavable linker between the first and second regions thereof. The linker, which is typically a polypeptide chain cleavable by a protease, or by other means suitable for effecting polypeptide cleavage, may be cleaved after production of the fusion protein in order to facilitate recovery of the desired polypeptide.

Moreover, in the event that the fusion protein comprises the chaperone fragment and the desired polypeptide as separate chains, held together otherwise than by a peptide bond, the cleavable linker may comprise an alternative cleavable site, such as a disulphide bond.

Preferably, the chaperone fragment is located N-terminal to the desired polypeptide in the fusion protein. Advantageously, the chaperone fragment itself forms the N-terminus of the fusion protein; however, it is envisaged that alternative N-termini may be included, such as to protect the chaperone fragment from degradation.

In the context of the present invention, although reference is made to both proteins and polypeptides, the terms are intended to be substantially interchangeable, with the exception that the term "protein" specifically includes multi-chain molecules comprising more than one polypeptide chain. The polypeptide chains may be held together by non-covalent or covalent means, wherein such covalent means do not include peptide linkages. For example, the chains may be held together by disulphide linkages.

The invention further comprises a nucleic acid encoding the fusion protein of the invention, and preferably the nucleic acid forms part of an expression vector comprising the nucleic acid operably linked to a promoter.

Nucleic acids, vectors and promoters are further described below.

The invention further comprises a host cell carrying the expression vector of the invention, and a method of preparing the fusion protein of the invention comprising (i) culturing the host cell under conditions which provide for the expression of the fusion protein from the expression vector within the host cell; and (ii) recovering the fusion protein from the cell.

In cases where the fusion protein further comprises a protease cleavable linker region between the first and second regions the method optionally further comprises cleaving the protein at the protease cleavable linker and recovering the second region.

Brief Description of the Figures

Figure 1 is a diagram showing plasmid pHGro, comprising an N-terminal histidine tag, the 191-345 fragment of GroEL, a thrombin cleavage site and a multiple cloning site.

Figure 2 shows an SDS-PAGE analysis of pHGro expression and purification systems. Molecular weight standards in the range 14,000 - 70,000 Daltons are loaded in lanes 5, 10 and 15 (Sigma #SDS-7 Dalton Mark VII-L™). Analysis of the sonication extracts shows that the Tenascin (lane 1), RNase HI (lane 6) and FKBP 12 (lane 11) fusion proteins are all over-expressed to a high level. Following a three hour incubation with Nickel affinity resin, all visible traces of the fusion protein are removed from the Tenascin (lane 2), RNase H1 (lane 7) and FKBP 12 (lane 12) sonication extracts. Tenascin, RNase H1 and FKBP 12 fusion proteins are released into the elution buffers (lanes 3,8 and 13 respectively). Thrombin successfully removes the GroEL fragment from Tenascin (lane 4), RNase H1 (lane 9) and FKBP 12 (lane 14).

Detailed Description of the Invention.

A: First region.

The first region of the fusion protein of the invention may comprise any natural or synthetic chaperone fragment.

Chaperone fragments suitable for use in the present invention are described, for example, in International patent application WO98/13496, the disclosure of which is incorporated herein by reference.

chaperone polypeptide having an amino acid sequence selected from at least amino acid residues 230-271 but no more than residues 150-455 or 151-456 of a GroEL sequence substantially as shown in SEQ. ID. No. 1, or a corresponding sequence of a substantially homologous chaperone polypeptide, or a modified, mutated or variant thereof having chaperone activity.

The sequence of GroEL is available in the art, as set forth above, and from academic databases; however, GroEL fragments which conform to the database sequence are inoperative. Specifically, the database contains a sequence in which positions 262 and

267 are occupied by Alanine and Isoleucine respectively. Fragments incorporating one or both of these residues at these positions are inoperative and unable to promote the folding of polypeptides. The invention, instead, relates to a GroEL polypeptide in which at least one of positions 262 and 267 is occupied by Leucine and Methionine respectively.

The amino acid sequence is preferably selected from at least amino acid residues 193-335, preferably 193-337, more preferably 191-345, even more preferably 191-376 but no more than residues 151-455. The invention therefore includes polypeptides being GroEL amino acid residues 230-271, 230-272 ...*et seq*... 230-455 and in like manner residues 230-271, 229-271 ...*et seq*... 151-271. Also, residues 230-271, 229-272 ...*et seq*... 151-351, 151-352 ...*et seq*... 151-455. All amino acid sequences of 42 or more residues comprising at least contiguous residues 230-271 and not exceeding 151-455 are within the scope of this aspect of the invention e.g. 171-423 or 166-406.

In a highly preferred aspect, the invention provides fragments selected from the group consisting of residues 191-375, 191-345 and 193-335.

There are four key properties that may characterise a protein as a molecular chaperone (1) suppression of aggregation during protein folding; (2) suppression of aggregation during protein unfolding; (3) influence on the yield and kinetics of folding; and (4) effects exerted at near stoichiometric levels. Fragments which possess these activities are suitable for use in the present invention.

Chaperone activity may be determined in practice by an ability to refold cyclophilin A but other suitable proteins such as glucosamine-6-phosphate deaminase or a mutant form of indoleglycerol phosphate synthase (IGPS) (amino acid residues 49-252) may be used. A rhodanese refolding assay may also be used. Details of suitable refolding assays are described in more detail in the specific examples provided hereinafter.

Preferably, the chaperone activity is determined by the refolding of cyclophilin A. More preferably, 8M urea denatured cyclophilin A (100μM) is diluted into 100mM potassium phosphate buffer pH7.0, 10mM DTT to a final concentration of 1μM and then contacted

with at least 1μM of said polypeptide at 25°C for at least 5 min, the resultant cyclophilin A activity being assayed by the method of Fischer G *et al* (1984) Biomed Biochim Acta 43: 1101-1111.

The polypeptide is preferably an hsp60 polypeptide, preferably a GroEL polypeptide.

A preferred polypeptide has the amino acid sequence 191-345 or 191-376, more preferably 193-335 or 191-337 of GroEL, or the equivalent residues of substantially homologous chaperonins, or a modified, mutated or variant sequence thereof.

The polypeptide preferably has a molecular weight of less than 34kDa.

"Modifications" include chemically modified polypeptides for example. "Variants" include, for example, naturally occurring variants of the kind to be found amongst a population of hsp60 chaperonin harbouring organisms/cells as well as naturally occurring polymorphisms or mutations. "Mutations" may also be introduced artificially by processes of mutagenesis well known to a person skilled in the art.

In being "substantially homologous" peptides may have at least 50% amino acid sequence homology with the specified GroEL amino acid sequences, preferably at least 60% homology and more preferably 75% homology. Homology may of course also reside in the nucleotide sequences for the polypeptide which may be at least 50%, preferably at least 60% homologous and more preferably 75% homologous with the nucleotide sequence encoding the specified GroEL amino acid residues.

Where conservative substitutions are made they may be made by reference to the following table, where amino acids on the same block in the second column and preferably in the same line in the third column may be substituted for each other:

ALIPHATIC	Non-polar	G A P
		I L V
	Polar - uncharged	C S T M
		N Q
	Polar - charged	D E
		R K
AROMATIC		H F W Y
OTHER		N Q D E

Synthetic variants of naturally-occurring chaperone proteins may be made by standard recombinant DNA techniques. For example, site-directed mutagenesis may be used to introduce changes to the coding region of a DNA encoding a naturally-occurring coiled-coil protein. Where insertions are to be made, synthetic DNA encoding the insertion together with 5' and 3' flanking regions corresponding to the naturally-occurring sequence either side of the insertion site. The flanking regions will contain convenient restriction sites corresponding to sites in the naturally-occurring sequence so that the sequence may be cut with the appropriate enzyme(s) and the synthetic DNA ligated into the cut. The DNA is then expressed in accordance with the invention to make the encoded protein. These methods are only illustrative of the numerous standard techniques known in the art for manipulation of DNA sequences and other known techniques may also be used.

The hsp60 class of chaperonin proteins are generally homologous in structure and so there are therefore conserved or substantially homologous amino acid sequences between the members of the class. GroEL is just an example of an hsp60 chaperonin protein; other suitable proteins having an homologous apical domain may be followed.

A fusion protein according to the invention will comprise as small a chaperone fragment as is feasible. This can be especially important where in structural determination of proteins by NMR it is often necessary to carry out isotopic labelling with ^{15}N or ^{13}C. This is expensive and with a long fusion partner much of the incorporated radioactivity is removed if the carrier protein (e.g. GST in many cases) is cleaved off.

B: Second Region.

The second region of the fusion protein according to the invention may comprise any polypeptide sequence of interest which is not naturally associated with the first region. Usually this will mean that the sequence of interest will be found in nature encoded by a gene different from the gene encoding the first region. This may be determined easily by examining the sequences of the first and second regions against publicly available sequence databanks. The second region may be from the same species as the first region, or from a different species. It is also possible that the first and second regions are derived from portions of the same protein but are present in the fusion protein of the invention in a manner different from the natural protein sequence.

The fusion protein according to the invention may be of any size although in general the invention is particularly useful when the polypeptide sequence of interest is short, e.g. from 2 to 100 amino acids in length, preferably 2 to 50 or even 2 to 30 or 5 to 10 amino acids in size. However larger polypeptide sequences of interest, e.g. up 150, 200, 400 or 1000 amino acids are also contemplated. The invention is particularly advantageous for the preparation of small polypeptides which are currently difficult to manufacture by recombinant means. Examples of such polypeptides include fragments of chaperone proteins, metabolic enzymes, DNA and RNA binding proteins, antibodies, viral proteins, intrinsic membrane proteins (including transport proteins from mitochondria, seven-helix receptor molecules, T-cell receptors), and cytoskeletal complexes, antibody binding peptides, peptide hormones (and other biologically active peptides made by ribosomal synthesis), and small subunits from multi-subunit biological structures such as respiratory enzymes, the ATP synthase. In general, the invention is suitable for use with peptides of any dimension, but the advantageous properties thereof are best exploited with small polypeptides, for example from 2 to 50 amino acids in length, particularly from 2 to 20 amino acids in length, and preferably from 5 to 10 amino acids in length.

A particular advantage of the present invention is that peptides may be produced by recombinant DNA technology which are so short that they would previously have been

made by oligopeptide synthesis techniques. Thus it is possible to produce libraries of peptides, for example of mutants of biologically active peptides, which may be screened or otherwise analysed, cheaply and efficiently in recombinant expression systems, particularly bacterial expression systems.

C: Cleavable linker region.

Where the first and second regions are linked by a cleavable linker region this may be any region suitable for this purpose. Preferably, the cleavable linker region is a protease cleavable linker, although other linkers, cleavable for example by small molecules, may be used. These include Met-X sites, cleavable by cyanogen bromide, Asn-Gly, cleavable by hydroxylamine, Asp-Pro, cleavable by weak acid and Trp-X cleavable by, *inter alia*, NBS-skatole. Protease cleavage sites are preferred due to the milder cleavage conditions necessary and are found in, for example, factor Xa, thrombin and collagenase. Any of these may be used. The precise sequences are available in the art and the skilled person will have no difficulty in selecting a suitable cleavage site. By way of example, the protease cleavage region targeted by Factor Xa is I E G R. The protease cleavage region targeted by Enterokinase is D D D D K. The protease cleavage region targeted by Thrombin is L V P R G.

D. Nucleic acids.

The invention also provides nucleic acid encoding the fusion proteins of the invention. These may be constructed using standard recombinant DNA methodologies. The nucleic acid may be RNA or DNA and is preferably DNA. Where it is RNA, manipulations may be performed via cDNA intermediates. Generally, a nucleic acid sequence encoding the first region will be prepared and suitable restriction sites provided at the 5' and/or 3' ends. Conveniently the sequence is manipulated in a standard laboratory vector, such as a plasmid vector based on pBR322 or pUC19 (see below). Reference may be made to Molecular Cloning by Sambrook *et al.* (Cold Spring Harbor, 1989) or similar standard reference books for exact details of the appropriate techniques.

Nucleic acid encoding the second region may likewise be provided in a similar vector system. Sources of nucleic acid may be ascertained by reference to published literature or databanks such as Genbank.

Nucleic acid encoding the desired first or second region may be obtained from academic or commercial sources where such sources are willing to provide the material or by synthesising or cloning the appropriate sequence where only the sequence data are available. Generally this may be done by reference to literature sources which describe the cloning of the gene in question.

Alternatively, where limited sequence data are available or where it is desired to express a nucleic acid homologous or otherwise related to a known nucleic acid, exemplary nucleic acids can be characterised as those nucleotide sequences which hybridise to the nucleic acid sequences known in the art.

Stringency of hybridisation refers to conditions under which polynucleic acids hybrids are stable. Such conditions are evident to those of ordinary skill in the field. As known to those of skill in the art, the stability of hybrids is reflected in the melting temperature (Tm) of the hybrid which decreases approximately 1 to 1.5°C with every 1% decrease in sequence homology. In general, the stability of a hybrid is a function of sodium ion concentration and temperature. Typically, the hybridisation reaction is performed under conditions of higher stringency, followed by washes of varying stringency.

As used herein, high stringency refers to conditions that permit hybridisation of only those nucleic acid sequences that form stable hybrids in 1 M Na+ at 65-68 °C. High stringency conditions can be provided, for example, by hybridisation in an aqueous solution containing 6x SSC, 5x Denhardt's, 1 % SDS (sodium dodecyl sulphate), 0.1 Na+ pyrophosphate and 0.1 mg/ml denatured salmon sperm DNA as non specific competitor. Following hybridisation, high stringency washing may be done in several steps, with a final wash (about 30 min) at the hybridisation temperature in 0.2 - 0.1x SSC, 0.1 % SDS.

Moderate stringency refers to conditions equivalent to hybridisation in the above described solution but at about 60-62°C. In that case the final wash is performed at the hybridisation temperature in 1x SSC, 0.1 % SDS.

Low stringency refers to conditions equivalent to hybridisation in the above described solution at about 50-52°C. In that case, the final wash is performed at the hybridisation temperature in 2x SSC, 0.1 % SDS.

It is understood that these conditions may be adapted and duplicated using a variety of buffers, e.g. formamide-based buffers, and temperatures. Denhardt's solution and SSC are well known to those of skill in the art as are other suitable hybridisation buffers (see, e.g. Sambrook, *et al.*, eds. (1989) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, New York or Ausubel, *et al.*, eds. (1990) Current Protocols in Molecular Biology, John Wiley & Sons, Inc.). Optimal hybridisation conditions have to be determined empirically, as the length and the GC content of the probe also play a role.

Given the guidance provided herein, nucleic acids suitable for forming the first or second region of a fusion protein according to the invention are obtainable according to methods well known in the art. For example, a DNA of the invention is obtainable by chemical synthesis, using polymerase chain reaction (PCR) or by screening a genomic library or a suitable cDNA library prepared from a source believed to possess the desired nucleic acid and to express it at a detectable level.

Chemical methods for synthesis of a nucleic acid of interest are known in the art and include triester, phosphite, phosphoramidite and H-phosphonate methods, PCR and other autoprimer methods as well as oligonucleotide synthesis on solid supports. These methods may be used if the entire nucleic acid sequence of the nucleic acid is known, or the sequence of the nucleic acid complementary to the coding strand is available. Alternatively, if the target amino acid sequence is known, one may infer potential nucleic acid sequences using known and preferred coding residues for each amino acid residue.

An alternative means to isolate the gene encoding the desired region of the fusion protein is to use PCR technology as described e.g. in section 14 of Sambrook *et al.*, 1989. This method requires the use of oligonucleotide probes that will hybridise to the desired nucleic acid. Strategies for selection of oligonucleotides are described below.

Libraries are screened with probes or analytical tools designed to identify the gene of interest or the protein encoded by it. For cDNA expression libraries suitable means include monoclonal or polyclonal antibodies that recognise and specifically bind to the desired protein; oligonucleotides of about 20 to 80 bases in length that encode known or suspected cDNA encoding the desired protein from the same or different species; and/or complementary or homologous cDNAs or fragments thereof that encode the same or a hybridising gene. Appropriate probes for screening genomic DNA libraries include, but are not limited to oligonucleotides, cDNAs or fragments thereof that encode the same or hybridising DNA; and/or homologous genomic DNAs or fragments thereof.

A nucleic acid encoding the desired protein may be isolated by screening suitable cDNA or genomic libraries under suitable hybridisation conditions with a probe.

As used herein, a probe is e.g. a single-stranded DNA or RNA that has a sequence of nucleotides that includes between 10 and 50, preferably between 15 and 30 and most preferably at least about 20 contiguous bases that are the same as (or the complement of) an equivalent or greater number of contiguous bases from a known or desired sequence. The nucleic acid sequences selected as probes should be of sufficient length and sufficiently unambiguous so that false positive results are minimised. The nucleotide sequences are usually based on conserved or highly homologous nucleotide sequences or regions of the desired protein. The nucleic acids used as probes may be degenerate at one or more positions. The use of degenerate oligonucleotides may be of particular importance where a library is screened from a species in which preferential codon usage in that species is not known.

Preferred regions from which to construct probes include 5' and/or 3' coding sequences, sequences predicted to encode ligand binding sites, and the like. For example, either the

full-length cDNA clone disclosed herein as SEQ. ID. No. 1 or fragments thereof can be used as probes, especially for isolating first region-encoding genes. Preferably, nucleic acid probes of the invention are labelled with suitable label means for ready detection upon hybridisation. For example, a suitable label means is a radiolabel. The preferred method of labelling a DNA fragment is by incorporating $\alpha^{32}P$ dATP with the Klenow fragment of DNA polymerase in a random priming reaction, as is well known in the art. Oligonucleotides are usually end-labelled with $\gamma^{32}P$-labelled ATP and polynucleotide kinase. However, other methods (e.g. non-radioactive) may also be used to label the fragment or oligonucleotide, including e.g. enzyme labelling, fluorescent labelling with suitable fluorophores and biotinylation.

After screening the library, e.g. with a portion of DNA including substantially the entire desired sequence or a suitable oligonucleotide based on a portion of said DNA, positive clones are identified by detecting a hybridisation signal; the identified clones are characterised by restriction enzyme mapping and/or DNA sequence analysis, and then examined to ascertain whether they include DNA encoding a complete polypeptide (i.e., if they include translation initiation and termination codons). If the selected clones are incomplete, they may be used to rescreen the same or a different library to obtain overlapping clones. If the library is genomic, then the overlapping clones may include exons and introns. If the library is a cDNA library, then the overlapping clones will include an open reading frame. In both instances, complete clones may be identified by comparison with the DNAs and deduced amino acid sequences provided herein.

It is envisaged that the nucleic acid of the invention can be readily modified by nucleotide substitution, nucleotide deletion, nucleotide insertion or inversion of a nucleotide stretch, and any combination thereof. Such mutants can be used e.g. to produce a mutant that has an amino acid sequence differing from the sequences as found in nature. Mutagenesis may be predetermined (site-specific) or random. A mutation which is not a silent mutation must not place sequences out of reading frames and preferably will not create complementary regions that could hybridise to produce secondary mRNA structure such as loops or hairpins.

The foregoing methods may, of course, be applied to the identification and modification or generation of sequences useful in any part of the fusion protein of the invention. In particular, the sequence of the IF_1 polypeptide coiled coil provided herein as SEQ. ID. No. 1, or suitable fragments thereof as discussed above, may be used as a probe for the identification of further suitable sequences.

The first or second region may also be manipulated to introduce an appropriate restriction enzyme site at the terminus which is to be linked to the nucleic acid encoding the first region via a corresponding restriction enzyme site. Desirably the sites will be either the same or at least have matching cohesive ends. Of course, the first and second regions may be joined by alternative means; for example, first region may be incorporated into primers used to isolate or replicate the second region.

Where a protease cleavable linker region is required, this maybe introduced into the linked first and second regions (e.g. into the restriction site linking the two) or introduced into one or the other prior to their combination.

E. Expression vectors and host cells.

The nucleic acid encoding a fusion protein according to the invention, or constituent part(s) thereof, can be incorporated into vectors for further manipulation. As used herein, vector (or plasmid) refers to discrete elements that are used to introduce heterologous DNA into cells for either expression or replication thereof. Selection and use of such vehicles are well within the skill of the artisan. Many vectors are available, and selection of appropriate vector will depend on the intended use of the vector, i.e. whether it is to be used for DNA amplification or for DNA expression, the size of the DNA to be inserted into the vector, and the host cell to be transformed with the vector. Each vector contains various components depending on its function (amplification of DNA or expression of DNA) and the host cell for which it is compatible. The vector components generally include, but are not limited to, one or more of the following: an origin of replication, one

or more marker genes, an enhancer element, a promoter, a transcription termination sequence and a signal sequence.

Both expression and cloning vectors generally contain nucleic acid sequences that enable the vector to replicate in one or more selected host cells. Typically in cloning vectors, these sequences enable the vector to replicate independently of the host chromosomal DNA, and includes origins of replication or autonomously replicating sequences. Such sequences are well known for a variety of bacteria, yeast and viruses. The origin of replication from the plasmid pBR322 is suitable for most Gram-negative bacteria, the 2μ plasmid origin is suitable for yeast, and various viral origins (e.g. SV 40, polyoma, adenovirus) are useful for cloning vectors in mammalian cells. Generally, the origin of replication component is not needed for mammalian expression vectors unless these are used in mammalian cells competent for high level DNA replication, such as COS cells.

Most expression vectors are shuttle vectors, i.e. they are capable of replication in at least one class of organisms but can be transfected into another organism for expression. For example, a vector is cloned in E. coli and then the same vector is transfected into yeast or mammalian cells even though it is not capable of replicating independently of the host cell chromosome. DNA may also be replicated by insertion into the host genome. However, the recovery of genomic DNA encoding the fusion protein of the invention is more complex than that of exogenously replicated vector because restriction enzyme digestion is required to excise the DNA. DNA can be amplified by PCR and be directly transfected into the host cells without any replication component.

Advantageously, an expression and cloning vector may contain a selection gene also referred to as selectable marker. This gene encodes a protein necessary for the survival or growth of transformed host cells grown in a selective culture medium. Host cells not transformed with the vector containing the selection gene will not survive in the culture medium. Typical selection genes encode proteins that confer resistance to antibiotics and other toxins, e.g. ampicillin, neomycin, methotrexate or tetracycline, complement auxotrophic deficiencies, or supply critical nutrients not available from complex media.

As to a selective gene marker appropriate for yeast, any marker gene can be used which facilitates the selection for transformants due to the phenotypic expression of the marker gene. Suitable markers for yeast are, for example, those conferring resistance to antibiotics G418, hygromycin or bleomycin, or provide for prototrophy in an auxotrophic yeast mutant, for example the URA3, LEU2, LYS2, TRP1, or HIS3 gene.

Since the replication of vectors is conveniently done in E. coli, an E. coli genetic marker and an E. coli origin of replication are advantageously included. These can be obtained from E. coli plasmids, such as pBR322, Bluescript© vector or a pUC plasmid, e.g. pUC18 or pUC19, which contain both E. coli replication origin and E. coli genetic marker conferring resistance to antibiotics, such as ampicillin.

Suitable selectable markers for mammalian cells are those that enable the identification of cells competent to take up vector nucleic acid, such as dihydrofolate reductase (DHFR, methotrexate resistance), thymidine kinase, or genes conferring resistance to G418 or hygromycin. The mammalian cell transformants are placed under selection pressure which only those transformants which have taken up and are expressing the marker are uniquely adapted to survive. In the case of a DHFR or glutamine synthase (GS) marker, selection pressure can be imposed by culturing the transformants under conditions in which the pressure is progressively increased, thereby leading to amplification (at its chromosomal integration site) of both the selection gene and the linked DNA that encodes the fusion protein. Amplification is the process by which genes in greater demand for the production of a protein critical for growth, together with closely associated genes which may encode a desired protein, are reiterated in tandem within the chromosomes of recombinant cells. Increased quantities of desired protein are usually synthesised from thus amplified DNA.

Expression and cloning vectors usually contain a promoter that is recognised by the host organism and is operably linked to the fusion-protein encoding nucleic acid. Such a promoter may be inducible or constitutive. The promoters are operably linked to DNA encoding the fusion protein by removing the promoter from the source DNA by restriction enzyme digestion and inserting the isolated promoter sequence into the vector. Both the

native promoter sequence of one of the constituents of the fusion protein and many heterologous promoters may be used to direct amplification and/or expression of the DNA. The term "operably linked" refers to a juxtaposition wherein the components described are in a relationship permitting them to function in their intended manner. A control sequence "operably linked" to a coding sequence is ligated in such a way that expression of the coding sequence is achieved under conditions compatible with the control sequences.

Promoters suitable for use with prokaryotic hosts include, for example, the β-lactamase and lactose promoter systems, alkaline phosphatase, the tryptophan (trp) promoter system and hybrid promoters such as the tac promoter. Their nucleotide sequences have been published, thereby enabling the skilled worker operably to ligate them to DNA encoding the fusion protein using linkers or adaptors to supply any required restriction sites. Promoters for use in bacterial systems will also generally contain a Shine-Delgarno sequence operably linked to the DNA encoding the fusion protein.

Preferred expression vectors are bacterial expression vectors which comprise a promoter of a bacteriophage such as phagex or T7 which is capable of functioning in the bacteria. In one of the most widely used expression systems, the nucleic acid encoding the fusion protein may be transcribed from the vector by T7 RNA polymerase (Studier et al, Methods in Enzymol. 185; 60-89, 1990). In the *E. coli* BL21(DE3) host strain, used in conjunction with pET vectors, the T7 RNA polymerase is produced from the λ-lysogen DE3 in the host bacterium, and its expression is under the control of the IPTG inducible lac UV5 promoter. This system has been employed successfully for over-production of many globular proteins, but in many other cases significant over-production cannot be achieved because of the toxicity of over-expression (Studier *et al.*,1990; George et al, J. Mol. Biol. 235; 424-435, 1994) . Alternatively the polymerase gene may be introduced on a lambda phage by infection with an int- phage such as the CE6 phage which is commercially available (Novagen, Madison, USA). other vectors include vectors containing the lambda PL promoter such as PLEX (Invitrogen, NL) , vectors containing the trc promoters such as pTrcHisXpressTm (Invitrogen) or pTrc99 (Pharmacia Biotech,

SE), or vectors containing the tac promoter such as pKK223-3 (Pharmacia Biotech) or PMAL (New England Biolabs, MA, USA).

Moreover, the fusion protein gene according to the invention may include a secretion sequence in order to facilitate secretion of the polypeptide from bacterial hosts, such that it will be produced as a soluble native peptide rather than in an inclusion body. The peptide may be recovered from the bacterial periplasmic space, or the culture medium, as appropriate.

Suitable promoting sequences for use with yeast hosts may be regulated or constitutive and are preferably derived from a highly expressed yeast gene, especially a Saccharomyces cerevisiae gene. Thus, the promoter of the TRP1 gene, the ADHI or ADHII gene, the acid phosphatase (PH05) gene, a promoter of the yeast mating pheromone genes coding for the a- or α-factor or a promoter derived from a gene encoding a glycolytic enzyme such as the promoter of the enolase, glyceraldehyde-3-phosphate dehydrogenase (GAP), 3-phospho glycerate kinase (PGK), hexokinase, pyruvate decarboxylase, phosphofructokinase, glucose-6-phosphate isomerase, 3-phosphoglycerate mutase, pyruvate kinase, triose phosphate isomerase, phosphoglucose isomerase or glucokinase genes, the *S. cerevisiae* GAL 4 gene, the *S. pombe* nmt 1 gene or a promoter from the TATA binding protein (TBP) gene can be used. Furthermore, it is possible to use hybrid promoters comprising upstream activation sequences (UAS) of one yeast gene and downstream promoter elements including a functional TATA box of another yeast gene, for example a hybrid promoter including the UAS(s) of the yeast PH05 gene and downstream promoter elements including a functional TATA box of the yeast GAP gene (PH05-GAP hybrid promoter). A suitable constitutive PHO5 promoter is e.g. a shortened acid phosphatase PH05 promoter devoid of the upstream regulatory elements (UAS) such as the PH05 (-173) promoter element starting at nucleotide -173 and ending at nucleotide -9 of the PH05 gene.

Fusion protein gene transcription from vectors in mammalian hosts may be controlled by promoters derived from the genomes of viruses such as polyoma virus, adenovirus, fowlpox virus, bovine papilloma virus, avian sarcoma virus, cytomegalovirus (CMV), a

retrovirus and Simian Virus 40 (SV40), from heterologous mammalian promoters such as the actin promoter or a very strong promoter, e.g. a ribosomal protein promoter, and from the promoter normally associated with the gene encoding a component of the fusion protein, provided such promoters are compatible with the host cell systems.

Transcription of a DNA encoding the fusion protein by higher eukaryotes may be increased by inserting an enhancer sequence into the vector. Enhancers are relatively orientation and position independent. Many enhancer sequences are known from mammalian genes (e.g. elastase and globin). However, typically one will employ an enhancer from a eukaryotic cell virus. Examples include the SV40 enhancer on the late side of the replication origin (bp 100-270) and the CMV early promoter enhancer. The enhancer may be spliced into the vector at a position 5' or 3' to the coding sequence, but is preferably located at a site 5' from the promoter.

Advantageously, a eukaryotic expression vector encoding the fusion protein may comprise a locus control region (LCR). LCRs are capable of directing high-level integration site independent expression of transgenes integrated into host cell chromatin, which is of importance especially where the fusion protein gene is to be expressed in the context of a permanently-transfected eukaryotic cell line in which chromosomal integration of the vector has occurred, in vectors designed for gene therapy applications or in transgenic animals.

An expression vector includes any vector capable of expressing nucleic acids that are operatively linked with regulatory sequences, such as promoter regions, that are capable of expression of such DNAs. Thus, an expression vector refers to a recombinant DNA or RNA construct, such as a plasmid, a phage, recombinant virus or other vector, that upon introduction into an appropriate host cell, results in expression of the cloned DNA. Appropriate expression vectors are well known to those with ordinary skill in the art and include those that are replicable in eukaryotic and/or prokaryotic cells and those that remain episomal or those which integrate into the host cell genome. For example, DNAs encoding the fusion protein according to the invention may be inserted into a vector

suitable for expression of cDNAs in mammalian cells, e.g. a CMV enhancer-based vector such as pEVRF (Matthias, *et al.*, (1989) NAR 17, 6418).

Particularly useful for practising the present invention are expression vectors that provide for the transient expression of DNA encoding the fusion protein in mammalian cells. Transient expression usually involves the use of an expression vector that is able to replicate efficiently in a host cell, such that the host cell accumulates many copies of the expression vector, and, in turn, synthesises high levels of fusion protein. For the purposes of the present invention, transient expression systems are useful e.g. for identifying fusion protein mutants, to identify potential phosphorylation sites, or to characterise functional domains of the protein.

Construction of vectors according to the invention employs conventional ligation techniques. Isolated plasmids or DNA fragments are cleaved, tailored, and religated in the form desired to generate the plasmids required. If desired, analysis to confirm correct sequences in the constructed plasmids is performed in a known fashion. Suitable methods for constructing expression vectors, preparing in vitro transcripts, introducing DNA into host cells, and performing analyses for assessing expression and function are known to those skilled in the art. Gene presence, amplification and/or expression may be measured in a sample directly, for example, by conventional Southern blotting, Northern blotting to quantitate the transcription of mRNA, dot blotting (DNA or RNA analysis), or in situ hybridisation, using an appropriately labelled probe based on a sequence provided herein. Those skilled in the art will readily envisage how these methods may be modified, if desired.

The invention moreover provides an expression vector comprising a first nucleic acid sequence encoding a polypeptide capable of forming a coiled coil structure operably linked to a promoter capable of expressing the first nucleic acid sequence in a host cell, and, linked to the nucleic acid sequence, a cloning site permitting the insertion of a second nucleic acid sequence such that it is capable of being expressed in fusion with the first nucleic acid sequence. Such a vector is a useful vehicle for expressing nucleic acids

encoding any desired polypeptide in the form of a fusion protein according to the invention.

A further embodiment of the invention provides host cells transformed or transfected with the vectors for the replication and expression of polynucleotides of the invention. The cells will be chosen to be compatible with the vector and may for example be bacterial, yeast, insect or mammalian.

Such host cells such as prokaryote, yeast and higher eukaryote cells may be used for replicating DNA and producing the fusion protein. Suitable prokaryotes include eubacteria, such as Gram-negative or Gram-positive organisms, such as *E. coli*, e.g. *E. coli* K-12 strains, DH5α and HB101, or Bacilli. Further hosts suitable for fusion protein encoding vectors include eukaryotic microbes such as filamentous fungi or yeast, e.g. Saccharomyces cerevisiae. Higher eukaryotic cells include insect and vertebrate cells, particularly mammalian cells. In recent years propagation of vertebrate cells in culture (tissue culture) has become a routine procedure. Examples of useful mammalian host cell lines are epithelial or fibroblastic cell lines such as Chinese hamster ovary (CHO) cells, NIH 3T3 cells, HeLa cells or 293T cells. The host cells referred to in this disclosure comprise cells in *in vitro* culture as well as cells that are within a host animal.

DNA may be stably incorporated into cells or may be transiently expressed using methods known in the art. Stably transfected mammalian cells may be prepared by transfecting cells with an expression vector having a selectable marker gene, and growing the transfected cells under conditions selective for cells expressing the marker gene. To prepare transient transfectants, mammalian cells are transfected with a reporter gene to monitor transfection efficiency.

To produce such stably or transiently transfected cells, the cells should be transfected with a sufficient amount of fusion protein-encoding nucleic acid to form the fusion protein. The precise amounts of DNA encoding the fusion protein may be empirically determined and optimised for a particular cell and assay.

Host cells are transfected or, preferably, transformed with the above-captioned expression or cloning vectors of this invention and cultured in conventional nutrient media modified as appropriate for inducing promoters, selecting transformants, or amplifying the genes encoding the desired sequences. Heterologous DNA may be introduced into host cells by any method known in the art, such as transfection with a vector encoding a heterologous DNA by the calcium phosphate coprecipitation technique or by electroporation. Numerous methods of transfection are known to the skilled worker in the field. Successful transfection is generally recognised when any indication of the operation of this vector occurs in the host cell. Transformation is achieved using standard techniques appropriate to the particular host cells used.

Incorporation of cloned DNA into a suitable expression vector, transfection of eukaryotic cells with a plasmid vector or a combination of plasmid vectors, each encoding one or more distinct genes or with linear DNA, and selection of transfected cells are well known in the art (see, e.g. Sambrook *et al.* (1989) Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory Press).

Transfected or transformed cells are cultured using media and culturing methods known in the art, preferably under conditions, whereby the fusion protein encoded by the DNA is expressed. The composition of suitable media is known to those in the art, so that they can be readily prepared. Suitable culturing media are also commercially available.

Preferred bacterial hosts which may be used in the method of the invention include B strains of *E. coli* such as BL21 or a K strain such as JM109. These strains are widely available in the art from academic and/or commercial sources. The B strains are deficient in the lon protease and other strains with this genotype may also be used. Preferably the strain should not be defective in recombination genes.

Most preferably the strain is BL21(DE3), as disclosed in Studier *et al.* (1990). Bacteria obtainable by selection for improved heterologous polypeptide expression, optionally cured of the original vector, may also be used as host cells in the present invention. Particular bacteria include *E. coli* C43 (DE3) (deposited at the European Collection of

Cell Cultures (ECCC), Salisbury, Wiltshire, UK on 4th July 1996 as B96070445); *E. coli* C0214(DE3) (deposited at the National Collections of Industrial and Marine Bacteria on 25th June 1997 as NCIMB 40884); *E. coli* DK8(DE3)S (deposited at the National Collections of Industrial and Marine Bacteria on 25th June 1997 as NCIMB 40885); or *E. coli* C41(DE3) (deposited at the ECCC on 4th July 1996 as B96070444). Such bacteria, when cured, provide a host for the expression of fusion proteins of the invention and are especially suitable for the expression of fusion proteins whose expression is toxic to bacteria.

F. Production of fusion proteins and their processing.

Host cells of the invention may be cultured under conditions in which expression of the fusion protein occurs. The fusion protein may be recovered by any suitable means, for example affinity chromatography or HPLC. Where small fusion proteins are involved HPLC is particularly suitable.

The fusion protein may be cleaved, e.g. using an appropriate protease, to provide the polypeptide sequence of interest and this sequence may be recovered from the resulting mixture of first and second regions of the fusion protein.

Alternatively the fusion protein may find application as such, for example as an immunogen where the coiled-coils form aggregates. This avoids the necessity for preparing immunogenic material form small proteins and peptides by coupling them by separate chemical reaction to a carrier protein such as key-hole limpet hemocyanin (KLH).

H. Use in NMR studies

Fusion proteins according to the invention possess an extremely small fusion partner. One advantage thereof is that the fusion proteins may be employed directly in an NMR experiment without the fusion partner interfering in the spectrum received.

NMR analysis may be performed according to techniques and methodology which are known in the art, for example as described in K. Würtrich, "NMR of Proteins and Nucleic Acids", Wiley, New York, 1986, incorporated herein by reference.

The present invention is illustrated with reference to the following examples.

Example 1

Preparation of a GroEL fusion vector

The polymerase chain reaction (PCR) is used to generate a DNA fragment containing a N-terminal histidine tag, the 191-345 fragment of GroEl, a thrombin cleavage site and a multiple cloning site. The 5'- flanking PCR primer is 5'- AGA CGG ACT GCC ATA TGC ATC ATC ATC ATC ATC ATG AAG GTA TGC AGT TCG ACC - 3'. The 3'- flanking primer is 5'- ATT GAC CCC AAG CTT CGA ATT CCA TGG TAC CAG CTG CAG ATG TCG AGC TCG GAT CCA CGC GGA ACC AGA CCA CGG CCC TGG ATT GCA GCT TCT TCA CCC -3'. The template for the PCR amplification is as described in Zahn *et al.*, (1996) PNAS (USA) 93:15024-15026. The resulting fragment is cloned into Nde I and Hind III digested PRSETA (Invitrogen) to create pHGro (see fig. 1).

A Fibronectin type III domain of human Tenascin and human FKBP 12 are sub-cloned into pHGro using BamH I and EcoR I. Residues 2-62 of *S. cerevisiae* RNase HI are amplified from genomic DNA by PCR and subcloned via *Bam*HI and *Eco*RI restriction sites into the pRSETa vector (Invitrogen), which also contains a fragment of the GroEL chaperone protein and a histidine tag. The sequences of the primers used for PCR amplification are as follows;

5' GCACCTAGGCGTTCCGTTCCCTTGAAGATGCGC (forward)

5'-GGGAATTCAGGAACTTCCATAGTTAGATGTAGTATTTGG (reverse).

The vector is used to transform *Escherichia coli* strain BL21(DE3)C41 (Novagen; Miroux, B., and Walker, J. E. (1996) *J. Mol. Biol.* **260**, 289 - 298) which is used for the expression tests in 2xTY medium. Transformants are obtained using a polyethylene glycol method (Chung, *et al.* (1989) *Proc. Natl. Acad. Sci. USA* **86**, 2172 - 2175).

Example 2

Expression and purification of fusion proteins

A 2 litre shake flask containing 0.25 litre of 2XTY medium plus ampicillin at 50 μg per ml is inoculated with 4 C41 colonies containing a pHGro fusion vector. The culture is grown at 28 °C in an orbital shaker at 200 rpm. At an Ab 6OO value of 0.3, expression is induced with isopropyl B-D-thiogalactoside (IPTG), using a 50 μM final concentration. The cells are harvested about 20 hours after induction by centrifugation and re-suspended in 200 ml of 20 mM sodium phosphate buffer pH 7.2 + 150 mM NaCl + 10 mM β-mercaptoethanol + PMSF (0.5 mM final concentration). The suspension is sonicated on power level 8 on a Misonix Inc. Model No. XL2020 sonicator, using 1 second pulses and 3 seconds cooling on ice for a total of 12 minutes and centrifuged at 15 k rpm for 30 minutes. The insoluble fraction is re-suspended in 100 ml of sonication buffer, re-sonicated and re-centrifuged.

Purification is performed in a batch-wise manner. The centrifuged protein solutions are combined and 10 ml of Ni^{2+} charged iminodiacetic acid resin (Sigma) is added. The solution is stirred for 3 hours at 4 °C and the resin washed 3 times with 50 ml of sonication buffer followed by 2 times with 50 ml of 50 mM Trizma base / Trizma HCl (Sigma) buffer pH 8.4 + 10 mM mercaptoethanol. Centrifugation is used to isolate the resin following each 1 minute wash. This process is repeated for each pHGro fusion. 50 ml of buffer containing 250 mM Imidazole is used to elute the fusion proteins from the resin. 50 mM Tris buffer pH 8.4 + 10 mM B-mercaptoethanol is used for human FKBPI2, pH 7.4 is used for Tenascin and RNase HI. It is necessary to include 150 mM NaCl during the elution of the RNase HI domain. 600 units of Thrombin (Sigma) is added to FKBP 12 and 50 units to the other two. After about 20 hours at room temperature the purifications are analysed by SDS-PAGE (Shagger, H., and Jagow, G. (1987) *Analytical Biochemistry* **166**, 368 - 379). Protein concentrations are estimated using Bio-Rad's Protein Assay solution, which is based on the Bradford dye-binding procedure. Bovine Serum Albumin is used to produce the calibration curve.

The test proteins are produced to an average of 400 mg per litre of culture, which is approximately 30% of the total soluble protein. All three fusion proteins behave in a typical manner during metal affinity chromatography, and thrombin removes the GroEL fragment successfully in each case (see FIG.2). Tenascin and RNase HI only require a small quantity of thrombin for the complete removal of GroEL. In the case of FKBP 12, a small amount of fusion protein remains after the treatment with thrombin. This has also been experienced with other FKBP 12 fusion proteins where thrombin has been used and is to be expected.

The 191-345 apical fragment of GroEL with a N-terminal histidine tag satisfies the criteria for a good fusion protein. It can be over-expressed to high levels as soluble fusion proteins in *E.coli*, it is small and can be purified easily using nickel affinity chromatography. Being monomeric, this expression system does not suffer from the problems associated with the expression of multimeric proteins with dimeric fusion proteins.

Example 3

NMR sample preparation

Uniform labelling of proteins with ^{15}N, or ^{15}N and ^{13}C is achieved by growing cells in minimal media containing $^{15}NH_4Cl$ or $^{13}C_6$-glucose as nitrogen and carbon sources respectively. A 10 % ^{13}C-labelled protein is produced by incorporating 10 % $^{13}C_6$-glucose and 90 % unlabelled glucose into the growth medium. Half litre cultures are grown at 28 °C, 250 rpm shaking, to an optical density of 0.2 AUs at 600 nm. Protein expression is induced for 16 h with 0.2 mM isopropyl-D-thiogalactoside and harvested cells are resuspended in 16 mM Na_2HPO_4, 4 mM $NaH_2PO_4.H_2O$, 150 mM NaCl and 10 mM β-mercaptoethanol. Cells are subject to two rounds of sonication and cell lysates are centrifuged at 17,000 r.p.m. for 30 min. The supernatant is applied to a nickel affinity column (Sigma) and the fusion protein eluted with 50 mM Tris-HCl pH 8.4, 150 mM NaCl, 10 mM β-mercaptoethanol and 250 mM imidazole. Thrombin digestion of the fusion protein, using 5 U thrombin per ml protein, released the RNase HI fragment from the GroEL tag fragment. This is carried out for 2 h at room temperature. The RNase HI is purified from the GroEL fragment using a Heparin HyperD column (Sigma) with a gradient of 1 M NaCl (0-100%) in 50 mM Tris-HCl pH 8.4 and 10 mM β-

mercaptoethanol. RNase HI containing fractions are dialysed overnight against 50 mM acetate buffer pH 3.6 and 5 mM DTT, and concentrated in an Amicon concentrator.

NMR samples contained approximately 2 mM protein in 50 mM acetate buffer pH 3.6 and 5 mM DTT, in either H_2O with 10 % D_2O or 100 % D_2O.

Claims

1. A fusion protein comprising:

a) a first region comprising a fragment of a chaperone polypeptide; and

b) a second region not naturally associated with the first region comprising a polypeptide sequence of interest.

2. A fusion protein according to claim 1 which further comprises a cleavable linker region between the first and second regions.

3. A fusion protein according to claim 1 or 2 wherein the first region is at or proximal to the N-terminus of the protein.

4. A fusion protein according to any one of the preceding claims wherein the polypeptide sequence of interest is from 2 to 250 amino acids in length.

5. A fusion protein according to any one of the preceding claims wherein the polypeptide sequence of interest is of eukaryotic origin.

6. A fusion protein according to any one of the preceding claims wherein the first region comprises a polypeptide selected from the group consisting of residues 191-375, 191-345 and 193-335 of *E. coli* GroEL

7. A nucleic acid encoding the fusion protein according to any one of the preceding claims.

8. An expression vector comprising the nucleic acid of claim 7 operably linked to a promoter.

9. An expression vector comprising a first nucleic acid sequence encoding a fragment of a chaperone polypeptide operably linked to a promoter capable of expressing

the first nucleic acid sequence in a host cell, and, linked to the nucleic acid sequence, a cloning site permitting the insertion of a second nucleic acid sequence such that it is capable of being expressed in fusion with the first nucleic acid sequence.

10. A host cell transformed with the expression vector of claim 8 or claim 9.

11. A method of preparing a fusion protein comprising:

(i) transforming a host cell according to claim 10, which method comprises culturing the host cell under conditions which provide for the expression of the fusion protein from the expression vector within the host cell; and

(ii) recovering the fusion protein.

12. A method according to claim 11 wherein the host cell is *E. coli*.

13. A method according to claim 12 wherein the expression vector comprises a bacteriophage T7 promoter.

14. A method according to any one of claims 11 to 13 wherein the fusion protein further comprises a protease cleavable linker region between the first and second regions and which method further comprises cleaving the protein at the protease cleavable linker and recovering the second region.

15. A polypeptide when prepared by the method of any one of claims 11 to 14.

16. Use of a polypeptide capable of forming a coiled coil structure as a fusion partner in the construction of a fusion protein.

17. Use according to claim 16, wherein the fusion protein is a fusion protein according to any one of claims 1 to 7.

18. Use of a fusion protein according to any one of claims 1 to 6 in NMR studies.

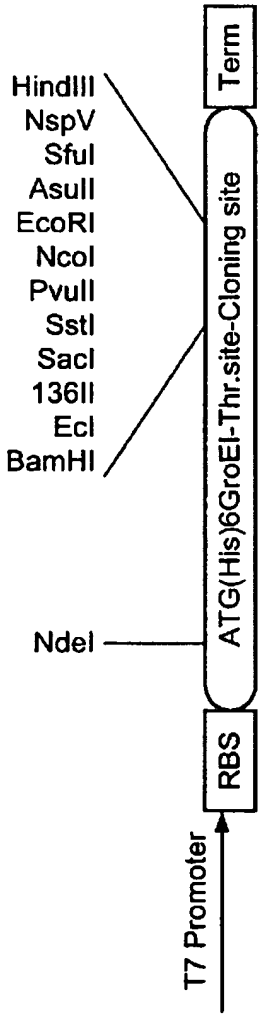

HindIII
NspV
SfuI
AsuII
EcoRI
NcoI
PvuII
SstI
SacI
136II
EcI
BamHI
NdeI
T7 Promoter
RBS
ATG(His)6GroEl-Thr.site-Cloning site
Term

FIG. 1

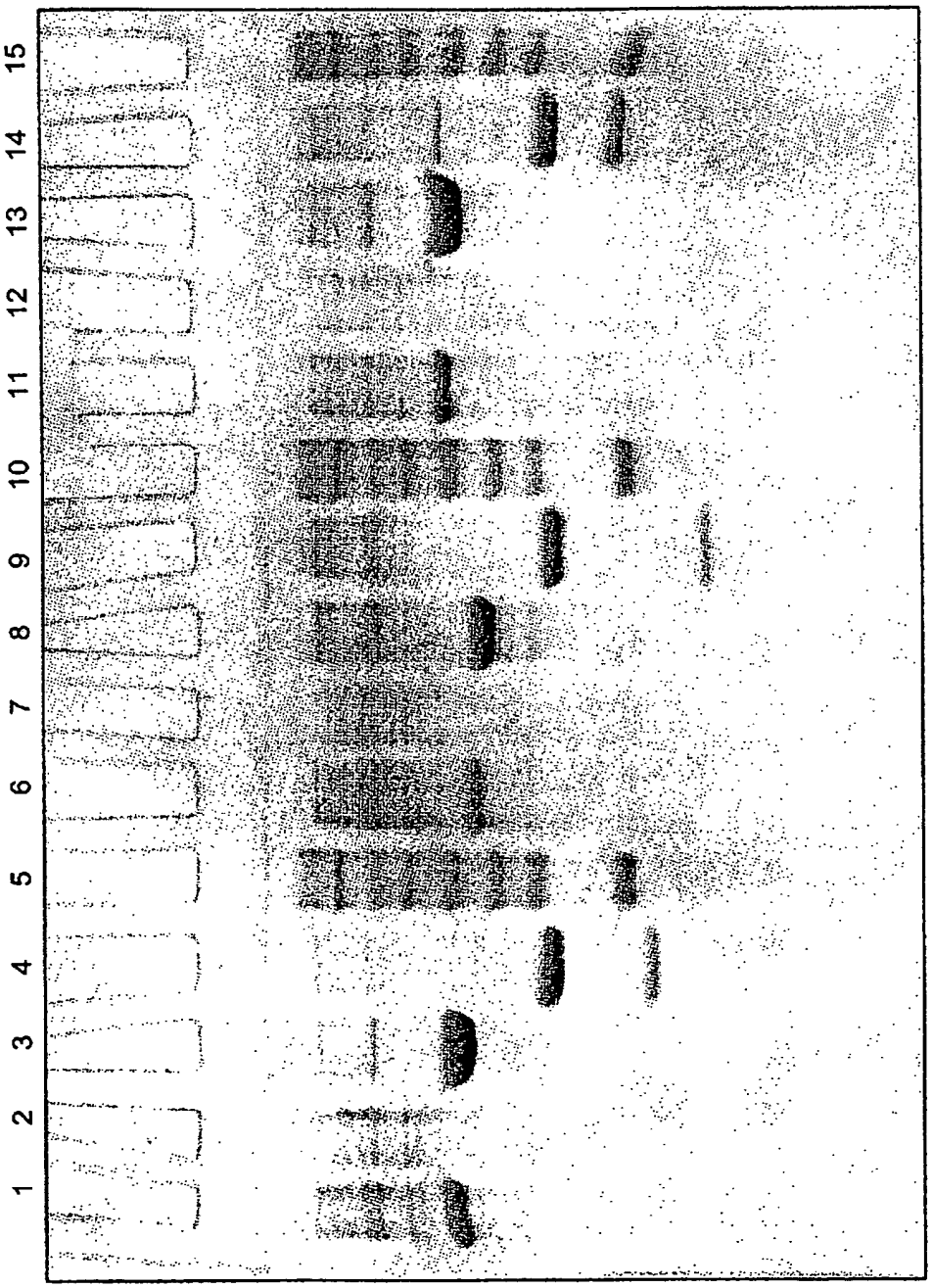


FIG. 2

INTERNATIONAL SEARCH REPORT

Int itional Application No
PCT/GB 00/01981

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12N15/62 C12N15/70 C12N1/21 C12P21/02 C07K14/47
C07K14/245 C07K1/113 //(C12N1/21,C12R1:19)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 C12N C12P C07K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
WPI Data, PAJ, CAB Data, STRAND, EPO-Internal, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	WO 93 13200 A (NOVONORDISK AS) 8 July 1993 (1993-07-08) page 4, line 25 -page 5, line 2; claim 20 page 10, line 26 - line 30	1,7,8, 10-12,15 2-6,9, 13,14, 16,17
X Y	WO 93 25681 A (UNIV NEW YORK) 23 December 1993 (1993-12-23) page 18, line 35 -page 19, line 1	1,7,8, 10-12,15 2-6,9, 13,14, 16,17

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
18 September 2000	29/09/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Hornig, H

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Ir. ational Application No
PCT/GB 00/01981

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
X	WO 97 18233 A (PHARMACIA & UPJOHN AB ;GUSTAFSSON JAN GUNNAR (SE); OEHMAN JOHAN (S) 22 May 1997 (1997-05-22)	1,7,8, 10-12,15
Y	the whole document	2-6,9, 13,14, 16,17
X	SAMUELSSON ELISABET ET AL: "Enhanced in vitro refolding of insulin-like growth factor I using a solubilizing fusion partner." BIOCHEMISTRY, vol. 33, no. 14, 1994, pages 4207-4211, XP002147681 ISSN: 0006-2960	1,7,8, 10-12,15
Y	the whole document	2-6,9, 13,14, 16,17
X	SAMUELSSON E ET AL: "FACILITATED IN VITRO REFOLDING OF HUMAN RECOMBINANT INSULIN-LIKE GROWTH FACTOR I USING A SOLUBILIZING FUSION PARTNER" BIO/TECHNOLOGY,US,NATURE PUBLISHING CO. NEW YORK, vol. 9, no. 4, 1 April 1991 (1991-04-01), pages 363-366, XP000572690 ISSN: 0733-222X	1,7,8, 10-12,15
Y	the whole document	2-6,9, 13,14, 16,17
Y	WO 98 13496 A (MEDICAL RES COUNCIL ;FERSHT ALAN ROY (GB); ZAHN RALPH (GB); ALTAMI) 2 April 1998 (1998-04-02) claims 1-50	2-6,9, 13,14, 16,17
Y	WO 98 24909 A (MEDICAL RES COUNCIL ;FERSHT ALAN ROY (GB); ZAHN RALPH (GB); ALTAMI) 11 June 1998 (1998-06-11) the whole document	2-6,9, 13,14, 16,17
Y	WO 99 02989 A (MEDICAL RES COUNCIL ;FERSHT ALAN (GB); CHATELLIER JEAN (GB)) 21 January 1999 (1999-01-21) the whole document	2-6,9, 13,14, 16,17
Y	WO 99 05163 A (MEDICAL RES COUNCIL ;FERSHT ALAN ROY (GB); ALTAMIRANO MYRIAM MARLE) 4 February 1999 (1999-02-04) the whole document	2-6,9, 13,14, 16,17

-/--

1

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 00/01981

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
Y	EP 0 774 512 A (IMANAKA TADAYUKI) 21 May 1997 (1997-05-21) the whole document	2-6,9, 13,14, 16,17
A	EP 0 650 975 A (NIPPON OIL CO LTD) 3 May 1995 (1995-05-03) cited in the application the whole document	
A	EP 0 412 465 A (HOECHST AG) 13 February 1991 (1991-02-13) the whole document	
P,X	WO 99 50302 A (TONGHUA GANTECH BIOTECHNOLOGY ;GAN ZHONGRU (CN)) 7 October 1999 (1999-10-07) the whole document	1-5,7-17

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/GB 00/01981

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9313200 A	08-07-1993	EP 0681609 A	15-11-1995
		FI 942907 A	17-06-1994
		JP 7504561 T	25-05-1995
		US 5681715 A	28-10-1997
WO 9325681 A	23-12-1993	NONE	
WO 9718233 A	22-05-1997	AU 705192 B	20-05-1999
		AU 7659696 A	05-06-1997
		CA 2236751 A	22-05-1997
		EP 0952981 A	03-11-1999
		JP 2000500023 T	11-01-2000
		NO 982155 A	12-05-1998
WO 9813496 A	02-04-1998	WO 9824909 A	11-06-1998
		AU 4467497 A	17-04-1998
		EP 0928336 A	14-07-1999
		AU 1036397 A	29-06-1998
WO 9824909 A	11-06-1998	AU 1036397 A	29-06-1998
		AU 4467497 A	17-04-1998
		EP 0928336 A	14-07-1999
		WO 9813496 A	02-04-1998
WO 9902989 A	21-01-1999	AU 8234698 A	08-02-1999
		EP 0995118 A	26-04-2000
WO 9905163 A	04-02-1999	AU 8547498 A	16-02-1999
		EP 0998485 A	10-05-2000
EP 0774512 A	21-05-1997	JP 9173078 A	08-07-1997
EP 0650975 A	03-05-1995	JP 7048398 A	21-02-1995
		US 5561221 A	01-10-1996
EP 0412465 A	13-02-1991	DE 3926103 A	14-02-1991
		AT 112286 T	15-10-1994
		DE 59007324 D	03-11-1994
		DK 412465 T	13-02-1995
		ES 2063874 T	16-01-1995
		IE 64771 B	06-09-1995
		PT 94934 A,B	18-04-1991
		US 5302518 A	12-04-1994
WO 9950302 A	07-10-1999	AU 6716498 A	18-10-1999